DOCUMENT RESUME

ED 298 912 IR 013 297

AUTHOR Baker, Eva L.
TITLE Evaluation Dimensions for Program Development and Improvement.
PUB DATE 84
NOTE 8p.; In: Instructional Development: The State of the Art, II; see IR 013 273. For volume I, see ED 292 444.
PUB TYPE Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Collection; *Evaluation Methods; *Evaluation Problems; *Evaluation Utilization; *Formative Evaluation; *Instructional Development; Program Evaluation; *Research Design

ABSTRACT

This chapter addresses the problems encountered in the formative evaluation of instructional development projects and the instructional development process. Three types of formative evaluation--component, convergent, and contextual--are distinguished, and the consequences of using the wrong type of evaluation in a particular situation or project are discussed. Various levels at which the evaluator may become involved in a formative evaluation are considered in the context of the accuracy of the data used in a formative evaluation. Recommendations are made for improvement based upon careful use of the three types of formative evaluation discussed in this chapter. Three figures are included. (10 references) (Author/EW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED298912

A. EVALUATION

Chapter 24

Evaluation Dimensions for Program Development and Improvement

Eva L. Baker

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Charles Dills
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Evaluation procedures are universally acknowledged as a critical component of any instructional development activity, yet it is often a poorly understood component of the system. This chapter provides information as to analysis of formative evaluation in terms of levels of function and relates types of formative evaluation to the development process. Component, convergent, and contextual approaches to formative evaluation are explored.

IR013297

INTRODUCTION

Every once in a while it is important to put some distance between common or cherished endeavors and one's perception of them in order to develop what is usually called perspective. People in the field of evaluation seem to take such reflective respites less often or less seriously than do others. Thus, the actors in evaluation seem suffused with a surreal sense of their own usefulness in the large context of educational development. The indisputable merit of the "good idea" of evaluation is rarely contrasted with the meager consequences of the practice of evaluation.

Those who happen upon the infrequency with which evaluation results are employed in educational development enterprises have explanations that are classifiable into two, usually exclusive, categories, political or technical. Evaluation is identified as a political act, and the general spector of "political" is thrown up as the encompassing explanation for the lack of real consequences, or technical improprieties or impurities in the conduct of the evaluation bear the burden for the lack of application. Both explanations derive from shared, and modest, expectations about the usefulness of evaluation. When politics is proffered as the explanation, typical reactions are knowing shrugs and "Isn't it the truth!" laments. When evaluation is seen to falter because of technical inadequacy, naturally we undertake efforts to become more adept technically. Both forms of response demonstrates that many evaluators function at a level they occasionally decry: their concern is more for the process (the doing of evaluations) than for the outcome (the consequences of evaluation activity).

Evaluators are not special in this regard. Given the complexities of our society, people are becoming accustomed to estimating satisfaction from their actual task, (for example, in social service the daily contact with clients, rather than from the long-term benefits of their actions, (how peoples' life-chances change). The long-term benefits appear to be just that: they are too "long-term" and too subject to the influence of the actions of others. Thus, we, in many areas, have given up "tracking" or expecting a linear relationship between our process and its desired or commonly-held outcomes.

Evaluators are not the only ones who have exhibited reliance upon their process rather than upon their outcomes; others have as well. For example, concern for the image of evaluation is

also becoming part of the routine rhetoric of full-time policy makers in legislative and executive roles. They know a good slogan when they hear one, and requirements to evaluate existing or developing programs abound in all areas of social action. Endorsing the adoption of an innovation without a formal affirmation of the anticipated utility of a future evaluation is unthinkable. Evaluation is seen as a device to assist public officials in their roles of protecting the public interest. Thus, a program that can demonstrate some actual evaluation component should be considered as reflecting the state-of-the-art at the political, if not at the technical, level.

This increasing dependence upon evaluation as a process separated from its planned consequences exacerbates the complexities of the act itself. For one thing, evaluators are regarded with skepticism. They have assumed tragi-comic social roles of "B" movie time/management consultants--very serious people who are sometimes not taken seriously. Evaluations are thought to expose error. And most normal people choose to defend themselves against public infamy. Although great differences are thought by evaluators to inhere in the common forms of evaluation, with the summative evaluation being a product or consumer report on competing programs, and formative evaluation being an evaluation to "help" the program developer create and improve the product, program developers are not as convinced. Their experience teaches them that the existence of negative data, even those generated in the early phases of program development for the most desirable intents, can be used against the program at some subsequent time. Thus, even within the "problem areas" of development, the image of evaluators as assisting in the improvement of programs has not been taken seriously; evaluators persistently make statements about what is wrong precursive to making recommendations for improvement.

Sensible program directors must find ways to contain the possible damage evaluation might wreak upon their commitment. They can choose to treat the evaluation arrangement as ceremonial, and thus limit its impact in that manner. Or they may reserve the option of discrediting the effort at some later point. The second option is fairly simple to do, given the idiosyncracy of evaluation and the lack of agreement about what constitutes satisfactory practice among those expert in evaluation. Any program director with a reasonable personal or organizational investment is normally wary when evaluators come to call. It is, of course, best when the wariness does not show.

So evaluators have at least two problems endemic in the area of evaluations with which they must contend at the outset. First, they must recognize that the increased tendency to endorse the process of evaluation occurs against a background of little concomitant evidence suggesting that evaluations are actually helpful or useful to various publics. (Evaluators can expect this anomaly to be noticed by others soon.) Second, program directors and members of other program-supporting constituencies have reasons to control the potential negative effects any evaluation, no matter what its guise, might generate.

If we accept that these two problems, exist, evaluators have the right to become somewhat shaky about their chosen line of work. Unfortunately, evaluators contribute to their own difficulties in numerous ways. One principal source of pain is that evaluators have not often attempted to clarify, for themselves, the roles and functions they will assume. Let's concentrate on formative evaluation[1], that type of evaluation which is principally regarded as contributing to the development and improvement of a program. Consider an example with the following set of limits. The evaluation team is working for the development group on a new course in remedial writing at the university level. That is, the evaluation team is separate from the development team. Thus, it is assumed that evaluation information will be generated for use by the program people, who will be making program modifications.

LEVELS OF FORMATIVE EVALUATION

Within formative evaluation, what might be some of the major and minor purposes addressed? At one level, the evaluator can be regarded as a safeguard against corruption, or unintended bias. In this role, the evaluator collects and summarizes information concerning the utility of a program, assuring that embarassing comments are not excised, or dismal data massaged in some way to present a rosier picture. Here the evaluator[2] acts to maximize the accuracy in the data collection, analyses and summary processes. The expertise required is primarily technical. While this use may seen trivial for persons of diverse talents, it represents a common function of many evaluators at present.

A second, more elaborate role for a person/team working in a formative instructional-improvement mode might involve participation in the selection or development of instruments and techniques used to collect information. In our example, the evaluation team would assist the program staff in the clarification of outcomes, and would suggest alternative ways in which they might be operationalized in performance indicators. Choices among tests and measures might be articulated. For instance, the pros and cons of using an indirect measure of writing skill, such as a multiple-choice measure, might be contrasted with the costs and benefits of collecting writing samples and using alternative scoring procedures. Here the evaluators might make comments, or even have the responsibility for scheduling of occasions for data collection and for timing the provision of feedback to the development staff.

At another level, evaluators may also suggest indicators other than those clearly appropriate to the stated goals of the project. In this expansion,

they may identify ways of assessing implicit goals, asking questions such as: "Is the course a relatively pleasant experience?" "Does it increase feelings of competence as well as the competency itself?" Further, the extent to which the teachers involved directly in the instruction find the format, content and organization of the program appropriate might also be assessed. The evaluator might suggest the use of observation, interview and questionnaire techniques for collecting this broader band of information. Occasionally, within formative evaluation, the development schedule is leisurely and the evaluator could opt to address longer-term outcomes of the program. For instance, opportunity to collect information on students' persistance at the unviersity, or performance in subsequent classes where writing might or might not be a major requirement, may also be recommended as such long-range indicators. At this third level, the evaluator generates a data base as relevant as possible to the range of information that the program developers can process. The evaluator may expand laterally the kind of data collected, and in doing so, gains considerable flexibility regarding how to weight multiple indicators and how to interpret patterns of support and conflict in the information gathered. The evaluator is still a technician, but must exhibit a good deal of flexibility and diversity in approach. A single-technique person will not do well in this role.

A fourth level of involvement requires much more range. The evaluation team may be expected to make recommendations concerning ways in which to improve activities and their organization, concerning alternatives to present program goals, and, in effect, to become substantively involved with plans for program modification. This is a role in which many evaluators may be comfortable, but it is one which requires not only common technical skills (involving measurement, statistical design, analysis, etc.), but also requires both a logical mind trained in pedagogy and an intimate knowledge of the subject area under consideration. The levels of function in formative evaluation are summarized in Figure 24-1.

When in the program Planner role, formative evaluators may need to tiptoe between the lines of careful analysis and of second-guessing when recommendations are made for program modification. The evaluator must demonstrate that such changes are the best that can be inferred from the data, should indicate other alternatives which have been disregarded, and is expected to use the full range of information collected to explain his or her recommendations. Evaluators rarely attend to their tasks with this level of detail. In those situations where this fourth level is expected, that is, where program-modification decisions are part of the anticipated report, evaluators are often both casual and speculative. And thus, this level of evaluation must be approached with great caution. Is the difficulty in this role simply a matter that, as usual, power (in this case, to change a program) corrupts? Is it simply a matter of finding a competent team which embodies the full range of skills required or producing substantive recommendations for the program? If the impact of the evaluation is to be more a matter of intent and less serendipity, another dimension must be explored. We must attend to what is a legitimate and proper function of the evaluator (as opposed to the creater) of educational programs.

EVALUATION FUNCTIONS

One way of trying to understand the evaluator's functions in the broadest sense is to look at them at a relatively high level of abstraction. Consider as an illustration the problem of understanding and verifying reality in the physical sciences. When techniques were few, and physics was the product of the relatively casual observation of and thought about natural events, attention was focused upon why something occurred. Philosophers wanted to understand. Analysis was under-

Figure 24-1. Levels of function in formative evaluation.

Level	Function	Role
1	Data collection, verification analysis report	Assistant to developer
2	Selection of measure + above	Assistant to developer
3	Generation of indicators, + above. Interpretation of data conflict	Collaborative
4	Recommendations for program revision, hypothesis generator + above.	Program planner

taken to develop casual links, albeit with little if any technical support. Aristotle wanted to know why the world and the heavens related as they did. As people developed more sophisticated techniques, they became interested in quite another level of analysis. They were concerned with relationships and processes. Newton and the physicists who followed him were attentive to how events occurred. They posited relationships and watched to see if they were correct. They forsook the understanding of higher order in return for predictability. Within the revolutions which have occurred and are occurring in natural science in the twentieth century, expectations continue to change. Our technological process has increased, for instance, in the use of devices to magnify and extend our perception of reality beyond imagination. We are at once, therefore, engulfed by information. Our refined data-collection techniques have demonstrated that the Universe is exceedingly complicated and consists of an extraordinary number of components, each of which would take lifetimes to unravel. As the means for observing reality have grown through technology, the grandeur of our questions has become reduced. Einstein and those who followed him have shown themselves to be interested in what is happening, as a point of departure, rather than the more ambitious questions of how and why. Simply finding procedures for the adequate description of events is an acceptably important first step. Positing casual and predictable relationships are tasks that are simply too hard. In reflecting upon the transition among the questions of why, how and what in physics, it seems that the parallel in educational evaluation might also be articulated.

Formative evaluators who attempt to link data with their recommendations are functioning at a Newtonian level of analysis. They think they can understand how processes are related to outcomes. One might properly argue, however, that the educating of people is exceedingly complex and that the possible relationships among processes are similarly entangled. For example, under experimental conditions, it has been shown that people can improve program performance by using tryout data for revising instruction (Baker, 1970; Rosen, 1968, Gropper and Lumsdaine, 1961). Yet in these studies, the analysis of other interesting variables was not undertaken, for instance, practice effects might appear to be a reasonable alternative hypothesis (the practice gained by the program developer extending his or her experience through another program revision effort). The attempt to isolate causes is mechanical, in the Newtonian sense. The level of technology is sufficiently developed in evaluation, that is "why" questions are no longer considered of primary interest. Thus, evaluators in formative evaluation roles do well to concentrate upon producing careful descriptions using the full range of technologies at their disposal in order to help people understand what has happened. They might also "polish up" the same explanation when peole ask whether the act of evaluation is actually related to improved programs.

This section has attempted to treat dimensions important to the conduct of formative evaluation. Dimensions of importance include the expectation of utility from the evaluation, and the likelihood of obtaining cooperation from others who have commitments to the educational program under scrutiny. It also included a discussion of the roles which formative evaluators adopt, from the technician to the hypothesis generator recommending program modification. A focus upon good description was suggested, and the point was illustrated through the transmutation of the roles of physicists, from the search to understand the meaning of events, to the desire to explore casual relationships, to the present concern with describing well the events that occur.

A recipe for good formative evaluation is offered. The evaluator must function at a level above that of being a mere technician. Alternative sources of data should be generated, but the limits upon the data-collection enterprise must be determined by the dimensions of control and timing. Data should be collected for those elements over which the program has some control. (For instance, demographic information is probably unimportant unless the program would be available differentially to various student populations, or unless alternative forms of training on program implementation were planned for different user groups.) Timing most simply means that the data must be generated in time for the program developer to do something with the information. Frequently, elaborated data-collection and analysis routines overlap with development schedules. When program development is fixed by special budgets then the insistence upon timely data collection is necessary, and the extent of description must be limited.

If the evaluator could choose to focus on a level of analysis appropriate both to their roles and to the available technology, then the ultimate form of the evaluation in program development is determined by the elements, such as the type of program to be evaluated and the form of development through which the program is proceeding. It may be antiquated to continue to discuss formative evaluation as if it were a single process; it is true that basic conventions for formative evaluation exist. But how did the codified idea of formative evaluation develop? Evaluation strategies for program improvement evolved largely because of the importance of verification in early small-scale program development efforts. For example, a basic statement of technique was prepared by Markle (1967). She described the steps in formative evaluation as beginning with small numbers of students working through a segment of instruction, through large samples of learners and expanded sections of programs and ultimately to "validation" samples where the final program would be used in "natural" conditions. Such strategy was appropriate because of two facts. First, the pro-

grams which Markle and others has in mind were materials-intensive. That is, students were exposed to self-contained materials in which the developers' influence was intended to be concentrated. Instruction was not complex, at least in form of administration. Mediation of instruction by live interpreters, such as teachers, was only an occasional feature of the instruction. What was then known as "programmed instruction" was the basis for this orientation. Second, the fact that the developer, or at that time, more properly the behavioral psychologist, was directly responsible for the instruction transmitted to students and for setting the rules by which data were to be analyzed. For example, issues of stimulus control were paramount. Data were collected to determine precisely the location in the program where more cues should be added in order to control student in-program responses. Performance outcomes were analyzed specifically to determine the place in the program where failure was occurring. Characteristics other than the psychological requirements of instructional design were only vaguely important. This assertion can be verified by checking the great number of misinformation and content errors in early examples of programmed instruction. Usually, although not exclusively, the first instructional programs were directed to the development of verbal repetoires, skill learning and rudimentary concept formation. The manner in which the instruction was to be organized within the network of other offerings in the school day was usually not considered.

Even when the scope of interest changed from programmed instruction to instructional development, the attributes of rcommended formative evaluation procedures carried their behavioral psycho-history. In this manner, formative evaluation has been characterized as requiring, by some, one-shot case study designs (as described by Campbell and Stanley in Gage, 1963), iterative engineering cycles, focus on student outcomes, and sample sizes range from two or three to multiples of the appropriate implementation units, e.g., multiple classrooms of students. A review of formative evaluation was reported by Baker and Alkin, 1973. Even in recent renditions, this orientation is dominant. The focus for evaluating first is in developing information sets about instructional design properties and then in extending and testing those designs with the full range of students and implementation sites. Data elicited from the teacher in many formative evaluation studies is related to their roles as conceived by the developer, as reincarnate of the behavioral psychologist. In the desire to transmit communication which as is as free of noise as possible to the student, the teacher is seen as a mediator, or really the medium, in the same sense as slide/tapes are medium through which presentations are made and stimuli controlled. Data are to verify that the teacher conformed to expectations.

Formative evaluation of this sort predominates in the field. Yet, it is clear that it is pertinent to only those development strategies which focus on refining segments in small numbers and extending work to large elements of program and subjects. This kind of formative evaluation we will call COMPONENT formative evaluation because it is based on incremental stacking of program components.

Alternative strategies of development call for varied formative evaluation styles. Another type of development, a delivery system strategy, is very different from the additive, linear development just described. In a delivery system orientation, the first problem for the developer is to visualize the final form of program operation. Product formats, design, training, the full variety of elements to be manipulated in the range of desired implementation settings are the first order of business. The ways in which teachers, students, materials and other support systems cohere into an operating program is established after an overall, operational system is functioning, then more refined instructional design elements are substituted into this operating sysem. Rather than conduct product testing on the instructional design, more traditional forms of experimental research are conducted on instructional variables of interest, e.g., response mode or number of examples. The critical feature of this approach is that this research can proceed in relatively traditional uncomplicated ways because it is not its purpose to determine how the findings relate to conditions of use. The verification or the formative evaluation, proceeds in two ways. One strand addresses the development of the operating system; and a second research activity parallels this effort. Only at the point where the researched instruction is substituted into the product system is more traditional formative evaluation of instructional elements a serious focus. For that reason, it is termed CONVERGENT formative evaluation. This form of evaluation has not traditionally been the orientation subscribed to in instructional development. The progression is not from laboratory to natural setting as in the Component form above. Rather, concern for the natural setting begins at the outset. Only when moderate confidence is reached for both the operating system and the instructional design elements do the pieces converge. While this orientation may seem to oppose in many ways programmed instruction, it does preserve one of its basic strategies: backward chaining. In programmed instruction (Gilbert, 1962) suggests that the best way to treat a problem was to approximate the final solution, e.g., tying a shoe, and decompose the necessary behaviors from that point. The learner is intended to develop a overview of what the final outcome is planned to be. A delivery system approach to development is similarly closed rather than open-ended. It is also probably much more efficient. Both Component and Convergent forms of development call for commitment to data collections for evaluation and research purposes, and to program-

matic and secure periods in which to engage in the task.

Unfortunately, many "innovative" programs, especially those funded through large scale State and Federal efforts have neither a focused, programmatic effort behind them, nor a protected period for research and development. These are diffuse programs and they are public issues as soon as the legislation is signed. They appear to succeed or fail primarily because constituencies in support or in opposition to the program are effective. Evaluators are often sought to assist in consensical judgments or at least to provide guidance to responsible policy makers on whether these programs are working and how they might be improved. In these cases, program guidelines are usually very general, a necessary attribute to accommodate to the full range of implementation sites and to the dominant notion of local control of innovation. Immediately, the idea of finding identifiable attributes of a program, except on the grossest levels, evaporates. The program description may consist of platitudes plus findings, plus considerable but enormously varied forms of effort. Finding dependable outcomes measures is easily as difficult as finding the "program process" itself.

In this case, the evaluator recognizes that the program evolves, and understands that the signal-to-noise ratio will be difficult to determine. Further, political instincts will be exposed and sensitivity to evaluation is likely to be high. For these reasons, the previous strategies of Component or Convergent formative evaluation are inappropriate. What the evaluator must do is to gather good descriptive information on the full range of outcomes and processes available. He or she should not attempt to limit sample size, except for economic reasons, but rather should take pains that the sample is truly representative of the range of settings in which the program is expected to operate. Fine-grained analysis and speculations of hypotheses to explain any findings may be developed through case study efforts, which are used to: 1) verify broad based data collection, and 2) provide a narration for policy makers on what kinds of things are actually happening in a integrated sense. Regression analyses of process and outcome data may generate hypotheses for further refinement, either through a smaller scale targeted study, or for quick and dirty planned variation in successive years of funding. The evaluator recognizes the great force of the ecology of the situation, including political and sociological factors. In this regard, the formative evaluation is CONTEXTUAL in nature, for context is expected to create maximum effect. An example of Contextual formative evaluation is the Evaluation of the California Early Childhood Education Program (see Baker, 1976, Baker, Herman, and Yeh, 1977). Figure 24-2 represents a summary of this section.

SUMMARY

Formative evaluation procedures should be matched to the development process in which they are embedded. Attempts to use formative evaluation methods of the Component type for development efforts that are contextual in nature will result in mismatched expectations and outcomes. The evaluator, once again, may have allowed his or her commitment to or experience with one form of

Figure 24-2. Types of formative evaluation and development process.

Development-Features	Developer Control	Formative Evaluation Type	Evaluation Features
Laboratory to Natural setting; incremental	High	Component	Interactive cycles, expansion of sample sizes
Operating system embedded in natural setting; parallel research on instructional design integration of both systems	Medium	Convergent	Controlled experiment research--formative evaluation of operating system
Highly diffuse and political programs; broad based, local options. Natural setting throughout	Low	Context	Broad representative sampling, regression analysis of process and outcomes

evaluation to color findings. Evaluators must begin to acknowledge that some of the early conceptual distinctions drawn for evaluation activity only poorly serve the full range of roles that evaluations are expected to satisfy. Perhaps the recognition that there exists component, convergent and contextual approaches to formative evaluation will allow our responsible growth as contributors to the improvement of education.

NOTES

[1]Although the distinction between formative and summative evaluation, as identified by Scriven (1967), was exraordinarily helpful, summative evaluation seems most appropriate in the preparation of discrete and clearly identifiable programs, such as specific products or product systems. I find the notion of summative evaluation less useful in connection with the existence of broad-based, broadly funded "programs" such as ESEA Title One. Although evaluations in this case are undertaken with a "summative" intent, clearly the program would continue to exist in some form and thus all data are to be used formatively.

[2]I use the term "evaluator" throughout for convenience. I am usually refering to a function rather than a single living, breathing body. In the same loose way, I might term "developers" those people with responsibility for development. This prose convention should not convey the notion that the competencies implied by either development or by evaluation tasks should reside in single persons. I am indebted to Richard E. Schutz for a seven-year conversation on this point.

REFERENCES

Baker, E.L. The evaluation of the early childhood education program, in Evaluation of the California Early Childhood Education Program. Volume 1, UCLA Center for the Study of Evaluation, December 1976.

Baker, E.L. Generalizability of rules for empirical revision. AV Communication Review, 1970, 18(3), 300-305.

Baker, E.L. and Alkin, M.C. Formative evaluation of instructional development. AV Communication Review, Winter 1973, 21(4).

Baker, E.L., Herman, J. and Yeh, J.P. Evaluation of the California Early Childhood Education Review. UCLA Center for the Study of Evaluation, November 1977.

Campbell, D.T. and Stanley, J.C. Experimental and quasi-experimental designs for research on teaching, in Handbook of Research on Teaching, N.L. Gage, Editor, 1963, 5, 171-246.

Gilbert, T.F. Mathetics: The technology of education. Journal of Mathetics, Mathetics, I, January 1962, 171-246.

Gropper, G.L. and Lumsdaine, A.A. The use of student response to improve Televised Instruction: An overview, in Studies in Television Instruction. Report No. 7, Pittsburgh: American Institute for Research, 1961.

Markle, Susan M. Empirical testing of programs. Programmed Instruction, University of Chicago Press, 1967, V, 104-140.

Rosen, M.J. An Experimental Design for Comparing the Effects of Instructional Media Programming Procedures: Subjective vs. Objective Revision Procedures. Final report, Palo Alto, California: American Institutes for Research in Behavioral Sciences, 1968.

Scriven, M. The methodology of evaluation, in R. Tyler, R. Gagne, and M. Scriven (Eds.), Perspectives of Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation No. 1, Chicago: Rand McNally, 1967.